

# Digital Libraries & Information Retrieval

Mark E Frisse, M.D.  
Washington University School of Medicine  
St. Louis, Missouri

## INTRODUCTION

Information is critical to effective medical practice. Information can be either ephemeral (e.g., telephone calls, hallway conversations, some email) or archival. Archival information, traditionally recorded in the printed word, is the organizational memory that sustains effective teaching and medical practice. Increasingly, this archival information will be produced in digital form, will be deposited in digital "virtual" health care libraries and will be retrieved at the point at which health care professionals and patients make medical decisions.

To be used effectively, information must be organized. And information relevant to the task at hand must be separated from information that is considered irrelevant. Information scientists often measure the relevance of a retrieved set of documents in terms of recall and precision. Recall may be defined as the percentage of all relevant items in a collection retrieved in response to a query; precision is similarly defined as the percentage of items retrieved that are relevant to a query.

This notion is familiar to all clinicians because of the similarity between the concepts of recall and precision applied to document retrieval and the concepts of sensitivity and specificity applied to diagnostic tests. In both areas, the laws are immutable. As one increases the sensitivity of a diagnostic test for example (and hence identifying more "true positive" individuals who have the condition measured by the test) one also decreases the specificity (and hence labelling more "false positive" individuals as "positive" for the measured condition). Similarly, as one casts a wider net by "ORing" additional search term, one is more likely to retrieve additional items of interest, but one is even more likely to retrieve items not relevant to the question at hand; and as one adds an increasing number of constraints to a query (by "ANDing"

additional query terms), one gets fewer items, but the items recovered are more likely to be relevant.

Successful information products match the appropriate information seeking client with the appropriate resource using retrieval technology assuring high precision and recall. They also present not just the right *kind* of information but also the right *amount* of information. Broad markets (e.g., Medline) require mass distribution and generic interfaces while narrow markets (e.g., a medical expert system) require greater retrieval or presentation efficacy and tolerate more idiosyncratic interfaces. Material of broad scope (e.g., these Proceedings) must be presented in a manner that facilitates rapid browsing and perspicuous presentation; material of narrow scope (e.g., an ascii email message) can be presented in a more terse and informal manner.

Ease-of-use is critical for success. Individuals seeking highly specific information on a drug reaction will reject a system that provides them with a comprehensive discussion of the drug. Individuals seeking a broader understanding will be frustrated if they are presented with information of only a narrow scope. Proper "fit" between need and delivery is essential.

## HEALTH CARE INFORMATION

Current and widely used retrieval technologies ensure that the manner in which information is organized and presented to the reader can differ from the manner in which it is represented when originally published. When thinking of composition, one should consider a multilevel representation. At the lowest level are "atomic" publications. These include published scientific articles with primary data, e-mail messages, records of medical charges in a relational database, and outcomes data. At a higher level, one sees more comprehensive reviews and

abstractions of these primary publications; examples include textbook chapters, patient care guidelines, the Physician's Desk Reference, GenRx, and other compendia. At the top level one finds information targeted to very specific audiences; examples include the Washington University *Manual of Medical Therapeutics* (aimed at internists early in their training), newsletters on specific clinical topics, and other content that attempts to bring the scope of an encyclopedic review to a narrow audience desiring only terse "guideline" information.

Medline - the topic of many articles in these and previous Proceedings - is often used in an attempt to create a "virtual book" by direct query and retrieval of abstracts and citations of the lowest level of published medical literature. Medline's greatest strength is in its scope; most of the published articles and reviews are indexed and presented, with abstracts, shortly after publication. To the individual, however, Medline's greatest shortcoming sometimes seems to be this very comprehensive nature. There are no "quality filters" implicit in the Medline database and it is said that over 50% of the literature is never cited by anyone and over 80% of the literature is cited at most only by its own author. In addition, physicians know that in some areas of medicine one can find a published article supporting virtually any position one may wish to take on a subject. In this sense, it is akin to a body of law where all opinions - majority or dissenting - are invoked as the rule of law.

Medline also suffers because it lacks the threading capabilities necessary in argumentation; its flat, phone-book structure compiles out the sequence and relationship of one article to another. For this reason, most novices and students are better off waiting for a critical review of a subject tailored to their needs rather than attempting to judge the merits of the primary literature. Unfortunately, the public is to a large extent driven by the press announcements surrounding early dissemination of primary research results. A researcher knows that publication of a high profile finding (a gene locus for heart disease, for example) will lead to national press coverage. If the findings are controversial, consumers made aware of this problem will bombard their providers with requests for an opinion on the merits of a case long before a review can be written - indeed, even many weeks before the article is indexed by the National Library of Medicine for Medline. Often, patients will be aware of a primary scientific publication before their providers; and both will be

in a state of ignorance about its implications to their health care. Medline is being used increasingly by skeptical or dissatisfied consumers who desire approaches to their illness. Issues once the province of informed, scientific criticism will more quickly become the subject of at times uninformed public debate.

Herein lies an opportunity. As long as the National Library of Medicine's indexing schedules are impeded by the necessary steps of professional indexing using controlled vocabularies, institutions and individuals able to respond rapidly to new findings will be perceived as more informed and decisive by their clients. On the long-term, providers will benefit from the value-added services of publishers who summarize and "quality filter" the primary literature in ways that minimize their own cost of information seeking and ease their development of patient care guidelines. The former, rapid response approach can evolve from better use of information professionals, from current awareness searches and from rapid automatic indexing of the published literature. As an example, consider the press coverage of bone marrow transplantation for breast cancer. This is a costly procedure whose merit is still unproven. How does a health care system respond to the popular press articles personalizing the plight of the individual breast cancer patient facing certain death and the few scientific articles of dubious merit that claim a beneficial response from this radical and expensive treatment?

Digital Libraries and information retrieval technology can be evaluated by the extent to which they enable professionals, patients, and large provider organizations in their efforts to identify and rapidly respond to controversial findings published in the primary literature and rapidly disseminated through the popular press

### TIMELY CONSENSUS

The papers in these Proceedings advance the identification and use of information resources that either support a specific action or help develop new and more efficient patient care guidelines. They jump-start the slow and laborious method of the traditional print-based published literature. They advance the development of consensus-building systems appropriate for patients, physicians, nurses, administrators, and executives. Examples of how these tools may ultimately be used include:

- An article from a news wire reports that a drug used to treat hypertension has been associated with an increased incidence of sudden death. Physicians and pharmacists are bombarded with phone calls from patients taking this medication. The American Heart Association, the drug manufacturer, and large health care systems each quickly and independently create rapid response teams to examine the information, develop policies, and disseminate their opinions to their clients.

- Responding to anecdotal reports of complications following early discharge after childbirth, a health care plan increases the covered length of hospital stay, even though they realize that an aggressive home health care and telephone follow-up system can provide lower marginal costs and greater savings. Unfortunately, the health plan lacks the ability to integrate information concerning resources and options into their prenatal education program, and without an efficient means of educating patients and reassuring all parties involved, the expense of additional hospital days is unavoidable.

- Wide differences of opinion exist among family practitioners treating coronary artery disease. Each practitioner has been influenced to a great degree by the practices of their referring cardiologists. Practitioners from a number of firms and hospitals have recently been merged into a single health care system, and the differences in perceived quality, cost, and process are of concern to primary practitioners, cardiologists, patients, and administrators. The health care system and insurers must find ways of developing a consensus quickly.

### AN EVOLUTIONARY PATH

Medical Informatics must focus both on the "what" of health care service provision as well as the "how" of service provision. The "what" can be studied through the distribution of information about health care options. For the consumer of health care services in a managed care environment, "open

enrollment" is a period in which one is deluged with contradictory information about the costs and benefits of various health care plans. It is only a matter of time until health care plans distribute this information via the Internet in an effort to assist their clients in identifying care providers and describing costs and benefits.

The "how" of health care is described in published or proprietary clinical practice guidelines that represent a consensus opinion. The long-term success of a health care delivery organization will depend on the extent to which its members work together to provide a seamless plan of care over extended periods of a client's life. This seamless consensus view is a strategic corporate asset every bit as valuable as knowledge concerning a critical manufacturing process. Like a manufacturing process, the ability to develop a complex, highly coordinated and flexible plan in minimal time and at minimal cost is a key to sustained growth. These plans are archival knowledge resources deserving of the technologies described in these papers.

Systems that address the "what" and "how" of health care must address the issues of collaboration, organizational knowledge, and digital commerce. Initially, such systems are composed primarily of published clinical literature, published guidelines, and institution-specific cost data. They should use information to express the collaborations necessary for the efficient delivery of health care. They should be expandable to allow incorporation of financial and clinical simulation data, decision analysis, and the strategic goals of a specific health care provider. They should be built more as compilers than as executable programs. A properly constructed system would in turn create a system that could be used to monitor the execution of a collaborative plan and, over time, modify plans to incorporate the experience gained. Both the "what" and the "how" of health care is an iterative process driven by information. As long as patients and technologies exist, there will be ways of improving health care delivery. This is a great opportunity for medical informatics researchers.